图书情报工作
第65卷 第5期 2021年3月

大规模中国历代存世典籍知识图谱构建研究*

■ 欧阳剑[1,2] 梁珠芳[3] 任树怀[1]
[1] 上海外国语大学图书馆 上海 201620 [2] 上海外国语大学新闻传播学院 上海 201620
[3] 广西民族大学管理学院 南宁 530006

摘 要：[目的/意义]探索构建中国历代存世典籍知识图谱，以为研究者挖掘海量古籍书目数据背后隐藏的知识提供一站式平台，拓展古籍知识服务内涵，同时，大规模的典籍知识图谱也是机器智能的重要基础。[方法/过程]通过知识图谱技术对中国历代存世典籍进行知识组织，从需求层、模型层、应用层3部分构建一个典籍知识图谱框架模型，通过人机协作进行典籍数据抽取及多源数据融合，完成数据的整理，并对典籍知识图谱实体类型及属性、典籍知识图谱实体关系及类型进行分析与定义。[结果/结论]所构建的典籍知识图谱包含649 549种古籍实体、221 783位典籍责任者、1 498 383个古籍版本、13 960个地名节点，形成了一个立体、多维、多用途的古籍知识关联网络，对全球目前存世的主要中国历代典籍书目信息进行了较全面描述。

关键词：古籍 知识组织 知识图谱 人文研究 数字人文
分类号：G250
DOI：10.13266/j.issn.0252－3116.2021.05.013

1 引言

古籍目录是一笔宝贵财富，也是引导我们打开我国古代文化遗产宝库的一把钥匙[1]，古籍目录也是文献学的重要组成部分，随着古籍数字化大潮来临，古籍文献的载体开始由原来的实物形态转换成以电子为载体、可以检索的数字形态，古籍目录的开发与利用过程中也在尝试采用新的信息技术。我国的目录学历史悠久，传统目录学、文献学面临着前所未有的冲击与挑战，古籍数字化更新了传统文献学的概念和内涵[2]，新数字化环境下出现了数字文献学理念[3]，提出了建立数字目录学的要求。建立数字目录是保护和弘扬中华文明的需要，用户可以完整系统地了解中华文明发展的脉络，古籍目录的开发也迎合了新的文献整理与研究者的需要。典籍知识图谱是古籍数字化的重要组成部分，是为了适应新信息环境而设计的一种语义知识组织和服务模式，它通过对典籍知识结构进行描述、揭示和表达，实现古籍知识管理和知识发现的目标，满足不同用户对知识表达和知识呈现的不同需求，从而更加智能地反馈给用户其所需的典籍基本信息和扩展信息。典籍知识图谱成为中国古籍文献研究平台的重要组成部分，也为数字文献学的研究提供了基础数据，建立中国存世典籍知识图谱是古籍知识服务的基础，可为中国古籍文献、历史学、哲学和语言学等领域的研究人员提供有效的帮助。

2 典籍目录知识化开发与利用现状

典籍目录整理及索引是早期典籍开发的主要形式，《中国古籍总目》《中国古籍善本书目》是典型代表，随着计算机技术的发展，数字化典籍目录及索引成为研究的主要方向，不少学者做了有益的探索[4-7]，此外，国际上也有不少针对中华典籍的相关研究。在数字化典籍目录开发与利用中，典籍知识化是一项重要的内容，何琳等[8]、罗晨光等[9]提出了基于本体的古籍知识建设，开始语义化、知识化的尝试。2009年，北京大学与国家图书馆所开发的《中国历代典籍总目分析系统》做了有意义的尝试，开发了具有划时代意义的古籍文献目录知识服务系统[10]。《中国历代典籍总目

* 本文系国家社会科学基金项目"图书馆古籍文献的数字人文开发与应用模式研究"（项目编号：17XTQ003）研究成果之一。
作者简介：欧阳剑（ORCID：0000－0001－5867－2852），研究馆员，博士，E-mail：oyjjj@163.com；梁珠芳（ORCID：0000－0003－3187－7502），硕士研究生；任树怀（ORCID：0000－0003－4817 －407X），馆长，研究馆员，教授。
收稿日期：2020－08－12 修回日期：2020－12－21 本文起止页码：126－135 本文责任编辑：易飞

分析系统》与传统的古籍目录应用相比取得了非常大的进步,但其中也存在不少遗憾,其构建的典籍知识单元还有所欠缺,虽然具有责任者、责任行为、版本特征以及装帧特征等多种维度的相关性分析功能,但由于缺乏时间、空间及各元素之间的关联数据支持,比如关键的编撰者信息缺乏,而无法从更多维度进行分析,数据颗粒度也较大,难以进行更精细化的分析,使得分析应用功能有一定的局限性。因此,有必要通过扩充与古籍相关联的人物信息、时间、地名等知识,把不同类型、不同颗粒度的古籍文献内容关联、整合和集聚起来,建立古籍知识关联网络,实现古籍知识存储、编辑、标引、知识挖掘和知识发现等功能,满足古籍内容价值深度挖掘和再创造需求[11],以进一步发现古籍内在的隐含知识,使传统的古籍内容大大增值。

在典籍的"辨章学术、考镜源流"功能开发研究中也有不少学者做了尝试,宋登汉等利用 RDA 体系从规范、书目、馆藏三个层次来设计古籍版本资源的整体描述,以期在古籍版本资源的描述上实现考证知识聚类功能[12-13]。邓仲华等使用本体库的构建技术针对古籍版本知识的数据进行了类、属性以及实例的设计[14]。夏翠娟等则提出了"古籍循证"的概念[15]。"辨章学术、考镜源流"功能只是典籍知识价值开发的部分应用,典籍知识更多、更有价值的应用有待深入。

总的来说,目前阶段针对于典籍知识的研究与开发应用相对有限。数字人文研究理念的出现促进了人文学科与技术的融合,也引发了古籍文献数据库建设、开发思路的转变[16],给古籍目录应用与开发带来了新的契机。在人文学科研究逐步强调"科学化"转型的过程中,知识关联、定量分析与挖掘是古籍文献深度开发与利用的发展方向[17],为古籍文献知识的深度开发与利用提供了新的理念与独特的创造性思维。

3 典籍知识图谱框架构建

3.1 典籍知识化需求分析

典籍目录不仅是引导治学的门径,更是考证学术源流的重要材料[18-19],以古籍目录为核心的应用主要在古籍版本源流考证方面[20]。典籍目录与古代学术文化有着密切的关系,古籍目录集成了古代文人的典籍之大成,典籍目录提供时空背景下的著作、出版情况,提供一种典籍的流传线索,为人们提供了另一个观察古代文人的地理分布、组合与变迁的角度,在一定程度上反映了中国经济、文化发展与社会变迁等,通过典籍目录能够反映历代典籍的流传、存亡状况,从中可推衍中国古代学术流变,也能够反映历代思想文化和学术旨趣[21]。典籍目录中的编撰者信息则是研究编撰者之间学术和社会关系的重要线索,近年来在文学地理学的研究中典籍目录与编撰者也成为重要的研究对象,通过分析历代文学家的地理分布情况,了解中国古代文坛的变化及古代学术的发展沿革情况,已成为文学地理研究的重要依据与手段[22]。

从人文研究应用的维度来说,需要围绕古籍形成年代、编撰者籍贯、收藏地等时间及空间角度进行分析。大规模典籍知识图谱的构建在强大的知识关联性方面有助于研究者全面观察古籍版本及版式信息,了解古代学术的发展沿革情况,考察版本源流,理清流变脉络,使用算法在古籍知识网络上可计算编撰者之间的学术和社会关系,从更深层次挖掘出我国古代文化的发展与变迁;还能够通过在古籍文献中分析编撰者、编撰时间、编撰方式、版本特征等多种维度的相关性,进一步揭示古籍数据背后隐藏的丰富知识,突破传统的单一数据源统计分析的模式,通过本体知识或者规则推理技术可以获取数据中存在的隐含知识,通过链接分析则可发现实体间隐含的关系,通过不一致检测技术可发现古籍数据编目中的噪声及差异;古籍编撰者空间信息可视化分析功能,能为文学地理的空间环境分析提供新的研究方式,更重要的是可为典籍研究者提供基础数据服务。

在传统与现代结合、机遇与挑战并存的古籍数字化大潮中既要保留优良传统,又要适应数字化时代的发展潮流,典籍知识服务最终的立足点是用户服务,需要有一套完备的平台,为学者构建一站式古籍目录检索系统[23],帮助研究者进行大规模的古籍目录收集、整理,通过不断满足不同用户的各种需求,加快整个古籍文献在内容、技术、研究等方面的创新升级,构建各类古籍知识如版本、版式、时间、地理、人物、编撰方式等的知识库并提供知识图谱服务,同时提供各种大规模典籍统计、分析、数据挖掘、知识推理等服务,大规模的典籍知识图谱也是机器智能的基础。

3.2 典籍知识图谱框架

知识图谱是近年来知识组织领域的研究热点,是一种以语义网络为基础的新型海量知识管理和服务模式[24]。构建知识图谱的主要目的是获取大量的、计算机可读的知识,构成网状的知识结构,增强知识单元之间的关联,实现用户主题检索需求,从而真正实现语义

检索[25]。近年来,知识图谱也开始在人文领域的研究中得到应用,特别是博物馆的文物知识及非遗文化组织领域,拓宽了传统人文数据存储维度和数据展现方式,实现了高效稳定的知识管理。本研究使用知识图谱的方法构建中国历代存世典籍知识图谱,将分散的典籍数据进行关联组织与重构,展示典籍知识之间的关联关系,为面向知识的挖掘和计算奠定基础,帮助学者发现隐性知识。

典籍知识是由古籍编撰者信息、收藏地及各种古籍目录元数据所组成,通过对碎片化的典籍知识单元进行有效组合,最终形成系统化的典籍知识库。典籍知识图谱构建分为需求层、模型层、应用层三部分(见图 1)。从需求层来说,典籍知识图谱构建要以需求为导向,了解人文研究的需求,在统一系统平台中对研究对象的多个属性数据采用知识图谱的形式进行组织,形成一个新的、更能有效表示该研究对象的综合数据集或获得新的隐性知识,借助数字人文研究的时间、地点、关系 3 个常用研究维度进行分析与挖掘:①以时间为主线分析研究典籍演进的轨迹过程,反映古典文学学术理念的发展;②对研究对象从地理空间进行分析和解读,包括各种空间元素及其结构(组合)与功能;③以研究对象的属性数据为基础,分析作品、编撰者、版本等之间的关系与结构。因此,古籍作品、编撰者、时间年代及地理信息等是典籍知识的重要组成部分,可以为研究者提供古籍书目、时间、地理、人物、版本、责任方式等多个分析角度。

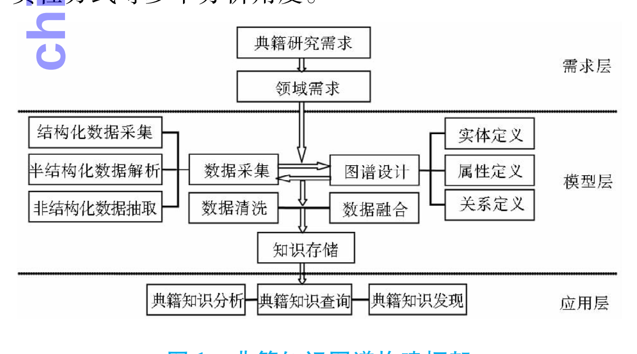


图 1 典籍知识图谱构建框架

基于数字人文研究的大规模典籍知识图谱构建主要是为了适应数字人文研究的需要,突破传统的应用及知识组织模式,充分挖掘古籍知识的最大价值。模型层是典籍知识图谱的核心,在典籍知识图谱构建中主要分为数据抽取、数据清洗、数据融合及知识图谱构建等几部分。古籍目录、古籍编撰者及地名信息等是典籍知识图谱的基础,古籍目录的主要属性有版本、版次、责任者、编撰方式、收藏地等,而编撰者主要属性有籍贯、所处朝代、官职等信息,地名主要包含责任者籍贯、藏书地等,这些信息来源既有结构化数据,也包括从散落的半结构化及非结构化信息中抽取的数据,因此,典籍信息抽取是典籍图谱构建的基础与关键,信息抽取主要通过信息发现、预处理和信息标注与提取,需要通过多种方法进行信息提取。数据融合也是典籍知识图谱构建的重要部分,负责对采集的信息进行清洗、合并及归一化处理。而图谱设计则是典籍知识图谱构建的实现环节,既包含概念模型的构建及实体、属性、关系的定义,也包含从典籍知识图谱概念到知识图谱的存储设计,最终形成典籍知识图谱知识服务系统。

应用层由典籍知识查询、知识分析、知识发现等特定应用服务模块和公用应用服务模块组成,每个应用服务模块提供特定的应用服务,典籍知识查询主要面向用户通用典籍知识查询服务,典籍知识分析则面向领域研究典籍分析,而知识发现则利用知识图谱的推理与计算优势辅助学者发现典籍知识中的隐性知识。

4 人机协作的典籍数据抽取及多源数据融合

4.1 典籍数据抽取及清洗

典籍知识图谱主要为典籍研究提供支撑,因此,典籍数据来源及数据准确性、真实性是知识图谱构建的关键,也是知识图谱的重要基础。本研究的典籍数据主要从国内外古籍书目网络数据库、全国普查数据、出版领域专业资料、通用领域的知识图谱、在线百科和万维网相关的网页抽取而来,主要通过数据抓取的方式采集相关网站典籍信息。抓取典籍目录数据相对来说比较简单,通过自主开发的采集软件并针对不同数据源设定相应采集规则即可完成对应数据源的典籍目录采集(见图 2)。

由于大部分典籍网站数据发布时对原始典籍编目数据进行了重组,大部分为如图 3 所示半结构化数据发布,因此需要对采集的数据进行不同程度的清洗,需要分别提取题名、朝代、编撰者、编撰方式等元数据,采用有监督(supervised)的方式在典籍目录提取过程中先对少量数据进行标注,然后进行机器学习,再使用学习模型对同类型或者符合特定关系的数据进行清洗,如:“杜工部草堂詩箋二十二卷(唐)杜甫撰(宋)魯訔編(宋)蔡夢弼箋(清)方功惠校訂”,经过数据清洗可成为如图 4 所示的结构化典籍数据。

欧阳剑，梁珠芳，任树怀. 大规模中国历代存世典籍知识图谱构建研究[J]. 图书情报工作,2021,65(5):126－135.

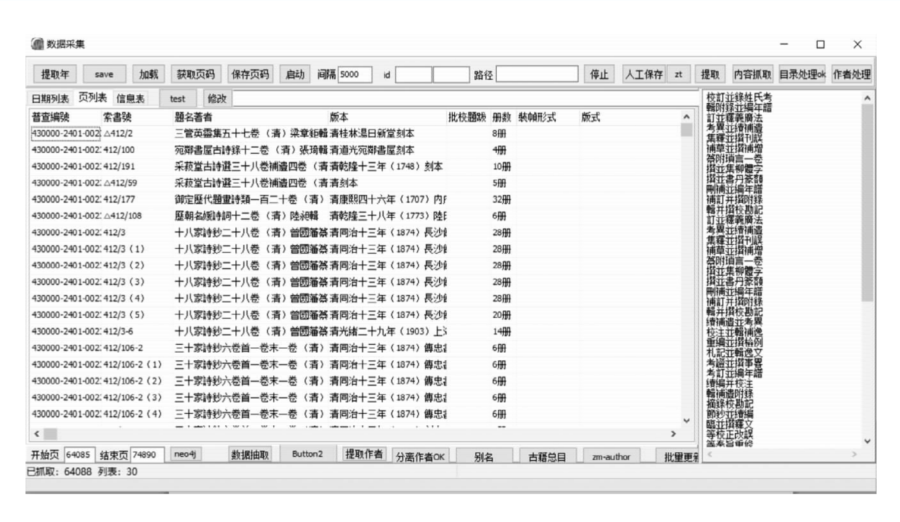


图 2　典籍数据采集

後漢書九十卷志註補三十卷（南朝宋）范曄撰（唐）李賢註

三國志六十五卷（晉）陳壽撰（南朝宋）裴松之注

晉書一百三十卷（唐）房玄齡等撰 晉書音義三卷（唐）何超撰

宋書一百卷（南朝梁）沈約撰

南齊書五十九卷（南朝梁）蕭子顯撰

梁書五十六卷（唐）姚思廉撰

图 3　半结构化典籍数据形式

杜工部草堂詩箋二十二卷

（唐）——杜甫——撰

（宋）——魯訔——编

（宋）——蔡夢弼——箋

（清）——方功惠——校訂

图 4　结构化典籍数据形式

相对而言，典籍编撰者信息是数据抽取的重点与难点。典籍编撰者从已经抽取的典籍数据中的编撰者项中进行提取，经去重后共有 221 783 位编撰者。在构建的典籍知识图谱中，编撰者实体包含朝代、生辰、字、号、别号、谥号、职业、籍贯、人物标签、代表作品、成就、官职等属性，主要通过 3 种方式获得：结构化数据信息抽取（如中国历代人物传记库（CBDB）、《中国历史人物辞典》等人名辞典）、半结构化数据信息抽取（在线百科类）、非结构化数据信息抽取（搜索网页），这三类数据中包含有丰富的典籍编撰者属性信息，如 CBDB 含有不少典籍编撰者信息，整个信息抽取流程见图 5。CBDB 及《中国历史人物辞典》等数据只有少部分能跟古籍的编撰者匹配，因此大部分数据需要通过在线百科与网络信息进行补充，分别用编撰者作为关键词进行检索，然后从百科类检索页面中抽取编撰者信息进行补充。百科类网站中的编撰者是一个个实体，每个实体的页面均围绕一个编撰者进行全方位的介绍，网页信息结构也比较固定，通过正则表达式配置相应的抽取模板即可进行编撰者信息抽取，百科类内容质量也比较高，因此，百科类网站成为许多知识图谱构建的首选。对于无法通过结构化、半结构化数据源抽取匹配的编撰者，则通过搜索引擎的方式查找典籍编撰者相关网页，由于返回的网页过多，需要构造一个二分类器来判断返回的网页是否是古籍文献编撰者介绍性网页，最后从该网页抽取编撰者信息。通过在线百科与搜索引擎的方式查找会存在典籍编撰者多义、编撰者信息属性不一致、多个对象属性值未分割、数值属性值格式不统一的情况，因此需要对数据进行清洗、同名排歧、数据对齐等处理，特别是与现代人物同名者比较多，在处理时通过简单的正则表达式去判别出生年、卒年的取值范围就能快速地判别出是否为古籍编撰者。

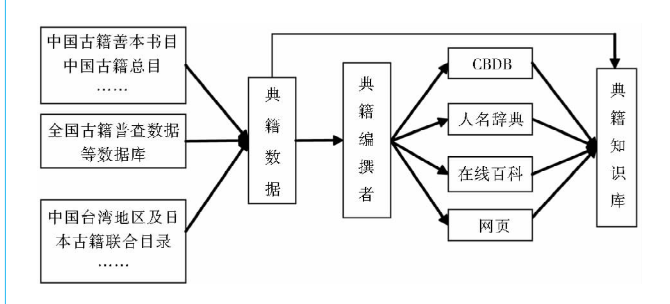


图 5　典籍知识库来源及抽取流程

4.2　多源典籍数据融合

异构知识资源的语义链接和集成是知识图谱的核心内容，需要研究异构数据的关联，将其转化成为具有丰富链接关系的知识网络。数据融合是对同一研究对象相关的多个属性数据采用一定的模式与方法，形成一个新的、更能有效表示该研究对象的综合数据集，将单一数据或不同类别的多源数据加以综合，消除多源信息之间可能存在的冗余和矛盾，加以互补，改善研究

对象信息提取的可靠性,提高数据的使用效率。典籍知识图谱也是一个数据高度融合的项目,典籍知识图谱包含多个异构的古籍目录、人物及地理数据,有来自不同图书馆的古籍编目数据、历史文献数据、书目资料、研究成果、网络数据等,需要将这些多源数据组织起来成为一个整体并进行合并,从而支撑各项研究对知识表达和知识呈现的需求,这一过程中的重要一步就是数据融合,如图 6 所示:

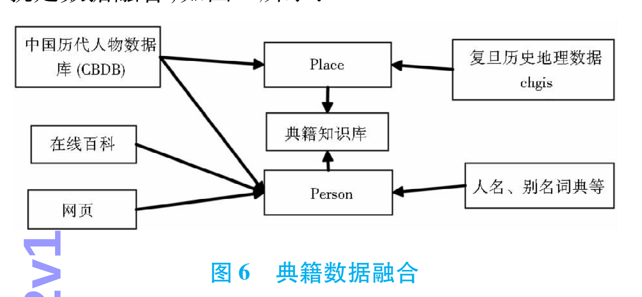


图 6　典籍数据融合

典籍数据融合主要是将古籍目录、人物及地名结构数据、半结构数据和非结构等不同形式的典籍数据进行融合,还要将不同来源的数据进行融合,在典籍知识构建中多源数据融合主要包含典籍目录、典籍编撰者、地名三类数据融合。典籍目录包含版本、藏书地、藏书数量等重要信息,由于典籍目录的来源不同,典籍目录的著录规则不统一,使得采集而来的原本属于同种典籍的信息存在一些差异,给典籍数据融合带来了难度。本研究主要通过题目 + 编撰者来确定是否属于同种典籍,当题目、编撰者一致时,则把它们归于同种典籍,本研究通过该方法将 250 万余典籍数据合并为 64.9 万余种不同古籍;典籍编撰者数据主要采用人名 + 朝代的方式进行编撰者关联,符合这个组合条件的数据被认为是同一编撰者,并提取相应编撰者的相关信息。此外,古籍目录的编撰者有的用别名或者号,如不加以处理也会造成不一致性,如:人名墨憨斋主人对应冯梦龙,需要通过实名与字号对应表进行映射;地名数据融合则主要需要处理不同朝代地名的变化,在不同的数据源中由于地名变化的差异也会导致无法对应,如古地名金城,即今甘肃兰州,唐称金城,又称金城郡等。数据的融合除了使用计算机进行自动处理外,必要的人工干预也是必不可少的(见图 7)。在对古籍目录、人物及地名数据等进行抽取并获取相关数据后,需在梳理和清理数据的基础上进行数据转换和建立结构化数据,实现数据整合和数据聚合,并建立基础数据集,同时利用多种数据源之间的冗余数据能对知识图谱的准确性进行合理的评估,冗余信息一方面可以提高知识点的可信度,另一方面也可以为后续人工编辑和校验提供参考依据,有利于消除古籍作品、版本、人名及地名的歧义。



图 7　典籍数据人工编辑和校验

5　典籍知识图谱构建实现及应用

典籍知识既有物质属性的一面,是记录知识内容的物质载体,同时,典籍也是物化了的思维、凝固的知识,因此典籍又具有精神属性的一面。典籍知识图谱的构建既要包含古籍的外在物质属性,又要包含古籍内在的隐含知识,并实现古籍文献特征的多维关联,达到典籍知识互通、共用。

欧阳剑，梁珠芳，任树怀. 大规模中国历代存世典籍知识图谱构建研究[J]. 图书情报工作,2021,65(5):126－135.

5.1 典籍知识图谱实体类型及属性

典籍知识图谱的核心是书目、版本、编撰者信息以及藏书地与编撰者籍贯相关的地名信息，根据典籍知识图谱使用场景确定采用书目、版本、责任者及地名 4 种实体类型。国际图书馆协会联合会编制的《书目记录的功能需求》为书目构建了一个概念模型，书目记录框架清晰定义了书目记录的实体、实体属性及实体间的各类关系，古籍书目借鉴《书目记录的功能需求》规范，在此基础上采用"作品－版本"的形式来进行表达，从概念(concept)上典籍知识图谱可归成 Work、Person、Version、Place 4 类(见图 8)，分别为作品、人物、版本、地名.《书目记录的功能需求》定义的作品的概念是抽象的，是独有的知识或艺术的创作，相对于古籍来说就是编撰者编撰的一种古籍书，特指具体的书目；版本则是一种书籍经过多次传抄、刻印或以其他方式而形成的各种不同本子，一个"作品"可以对应多种"版本"，一种"版本"可以有多个复本，有不同的收藏者；人物对应于作品的编撰者；地名是指对应版本藏书地与责任者籍贯，典籍知识图谱实体类型，见表 1。

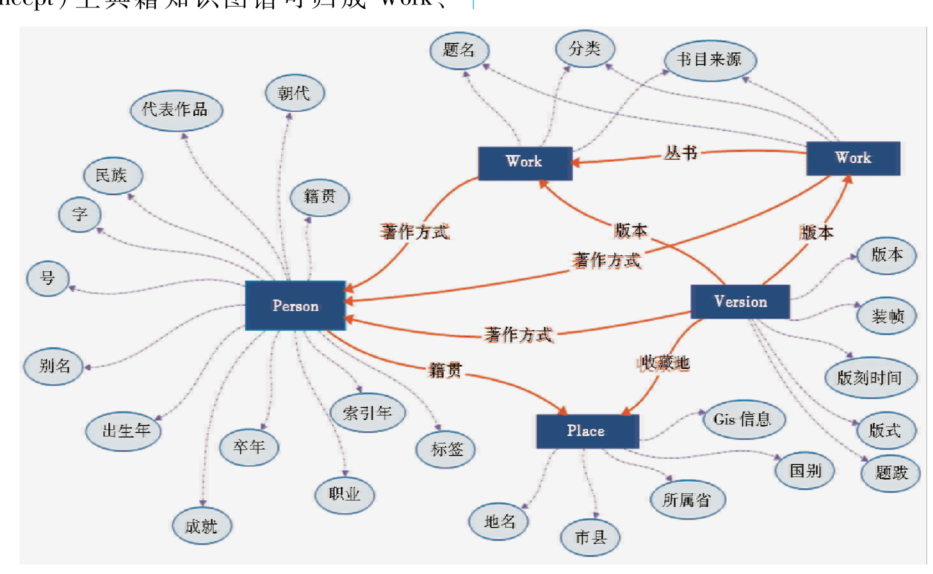


图 8 典籍知识概念

表 1 知识图谱实体类型

实体类型	中文含义	举例
Work	作品	《惜抱軒全集》《曾文正公手書日記》
Person	人物	曾國藩，姚鼐
Version	版本	清光緒刻古逸叢書本，清同治五年(1866)刻本
Place	地名	江苏泰兴，越州山阴(今浙江绍兴)

属性是典籍知识图谱对应实体(entity)的重要元素，每个作品都包含有题名、编撰者、编撰方式、所属分类、作品形成年代等属性，作品形成年代以编撰者所在年代为参考，编撰者朝代一般以卒年为断[26]，个别著者的朝代可参考其生平活动、成书年代及传统著录确定。而版本则可通过版刻时间和版本类型(版式)来描述，另外还包含该版本特有的编撰者、编撰方式。人物是典籍知识图谱中关联关系的重要对象，包含字、号、别名、出生年、卒年、代表作品、成就、标签、职业、籍贯、索引年(一般选卒年作为索引年，在卒年不详时则选其在文献中被提到的任职、活动事件等时间点为索引年)等属性。地名主要包含国别、省份、市县、名称、GIS 信息等属性。

5.2 典籍知识图谱实体关系类型

实体关系(entity relation)是指某一时间段内实体之间存在的关系。典籍知识图谱实体之间存在多种关系，主要有作品与编撰者、作品与版本、版本与收藏地、编撰者与籍贯等之间的关系。作品与编撰者之间存在编撰方式，通过对采集的古籍书目的编撰方式进行分析，发现其多达 2 千多种，为了便于研究统计分析，有必要对编撰方式归一化，根据《中华古籍总目编目规则》中的编撰方式要求"一般依正文卷端所题著录，原书所题性质相同或相近之著作方式，可适当归并而不尽据原题。"，除此之外作以下处理：①撰、著、述、学、拟、议等著作方式，统作为"撰"；②汇编整理前人著作者，统作为"编"；③辑录编次前人著作者，统称为"辑"；④抄录编次有关资料以成专书者，统作为"纂修"，除此之外则按依原题编撰方式建立作品与编撰者之间的关系。作品实体与版本实体则存在"作品版本"的关系，而版本与收藏地存在

“收藏于”的关系，编撰者与籍贯存在“属于”的关系，而有的作品是属于某个作品的一个子目，因此，除作品与编撰者存在多种关系之外，其他实体之间存在的关系则比较固定，典籍知识图谱实体关系及类型见表 2，实体之间通过关系连接形成了典籍知识概念图（见图 8）。

表 2　典籍知识图谱实体关系类型

实体	实体关系类型	中文含义	举例
责任者，作品	Way_of_works（纂…）	著作方式（纂）	<汪志伊，纂，荒政輯要九卷首一卷 >
作品，版本	version_is	版本	<荒政輯要九卷首一卷，版本，清道光十二年（1832）來鹿堂刻本>
版本，收藏地	held_in	收藏于	<清道光十二年（1832）來鹿堂刻本，收藏于，青海省圖書館>
责任者，籍贯	Born_in	出生于	<汪志伊，出生于，安徽桐城>
作品，作品	series_of_ books	丛书（子目）	<養正遺規二卷補編一卷，丛书，五種遺規>

5.3　典籍知识图谱实现

知识图谱以三元组模型表达“实体－属性”和属性值（statement），目前，知识图谱的存储主要为关联数据（linked data）、图数据库及关系数据库[27]，综合比较各知识图谱存储的优缺点，典籍知识图谱选择以图数据库 Neo4j 进行存储。在 Neo4j 中，知识单元由顶点（Vertex）、边（Edge）和属性（Property）组成的，其存储形式为三元组（S，P，O）数据，因此需要在典籍知识概念跟 Neo4j 存储之间建立映射，在 Neo4j 中节点类型跟典籍知识概念类对应，即节点分 Work、Person、Version、Place4 类实例，每个节点则对应相应的作品、编撰者、版本及地名实例，与数据库中数据的实体及属性建立对应关系，每个实例的属性通过属性名与属性值来标示，边对应实例之间的关系，边的属性则表示实体之间的关系类型（见表 3）。按照数据模型到数据库数据的对应规则转换对应数据成对应的数据集，将清洗、融合并归一化的数据导入 Neo4j 中，即可实现典籍知识图谱（见图 9）。

表 3　典籍知识与 Neo4j 对象

典籍知识	概念类	实例	实体之间关系	属性
Neo4j	节点类	顶点	边	属性名、属性值

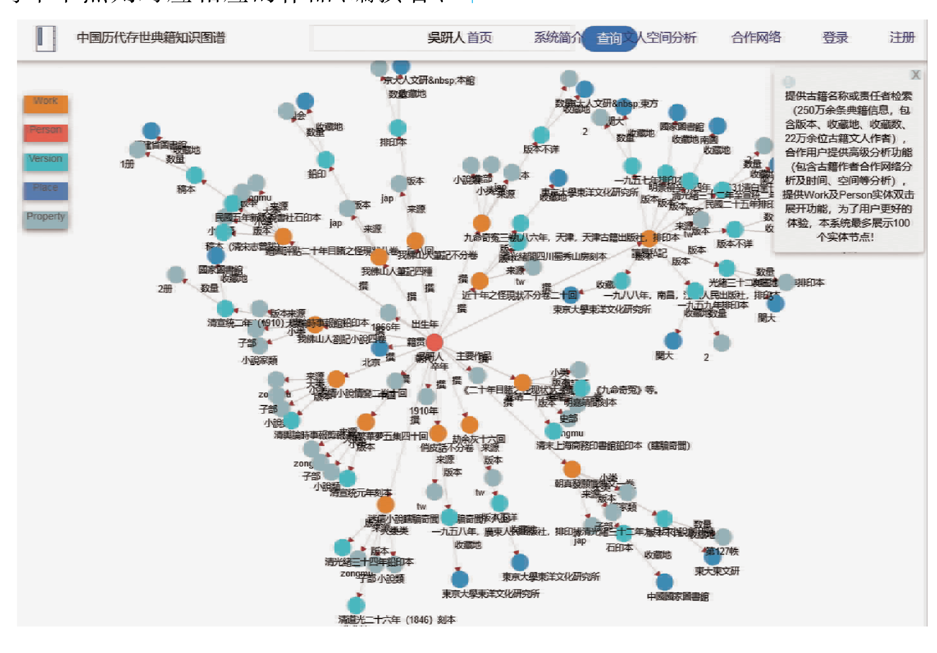


图 9　典籍知识图谱

最终形成的典籍知识图谱由来自于全球 743 家图书馆、科研院所等所藏的 250 万余中国历代存世典籍信息组成，其中包含古籍实体 649 549 种（Work 实例）、典籍责任者 221 783 位（Person 实例）、古籍版本 1498 383 个（Version 实例）、地名节点 13 960 个（Place 实例），这四类节点及其之间的关系构成一个庞大的典籍知识图谱，节点、属性及边等形成了一个立体、多维、多用途的古籍知识关联网络，对全球主要中国历代存

世典籍书目信息进行了较全面描述，为研究者挖掘海量古籍书目数据背后隐藏的知识提供了一站式平台，大大增强了古籍知识服务功能。

5.4 典籍知识图谱应用

知识图谱对实体关系具有表达能力强、对属性及结构可扩展性好、关联查询高效等优势，在对海量的多元异构信息进行建模时，图数据模型较关系模型、键值模型、文档模型等而言具有更直观的效果，更有利于使用者对数据结构和语义关系的理解。

典籍知识图谱作为一种基础性知识服务平台，首先，它能为普通大众提供基础典籍知识服务，通过简单的图谱可了解传统的典籍知识，增强文化传播效果。由于典籍数据来源的多样性，可以同时比较不同国家、地区同一种中国古籍的编目数据，发现古籍目录数据中的噪声及差异，为全国古籍普查工作提供编目参考，从而提高编目质量与工作效率。

其次，典籍知识图谱拓宽了典籍的应用范围，多维度的典籍知识图谱更为专业研究人员提供了深层知识挖掘和知识重组等高级服务，借助典籍知识图谱可分析典籍数据中存在的隐含知识，特别是在古籍版本对照、考镜源流方面，典籍知识图谱有着巨大的优势，能够依据相关图谱快速地了解版本特征以及装帧特征等多维度的相关性，还可以对典籍的成书年代、收藏地、收藏数量等进行分析(见图9)，获得定量的学术发展和研究重点的历史分布情况。

第三，典籍知识图谱也为相关人文研究提供了丰富的基础研究数据服务。典籍知识图谱从古籍实体、典籍编撰者、古籍版本、地名节点等不同知识维度组织，从多个角度对典籍进行了描述，为相关研究提供了强大的多维分析功能，借助于这些数据可以做更有深度的研究。比如，同一古籍的编撰者之间通常具有一定的关联关系，在版本、版次、印次和藏本层次上，古籍目录中的编撰者信息是研究编撰者之间学术和社会关系的重要线索，通过对这些编撰者著录信息进行定量分析可以获得较多的学术合作、学术传承、交游往来等关系，如通过合作网络的交互操作可发现与编撰者吴趼人存在直接或间接合作过的其他编撰者(见图 10)。文学与地理环境的关系是一个互动关系，中国历代文学家的地理分布格局分析是文学地理研究的重要内容，古籍文献的编撰者则是分析的主体，在传统研究中，从地理空间的视角研究文学，解析文本中的空间信息是一项繁杂的工作，本典籍知识图谱则包含有相关文人的多维数据，利用编撰者的籍贯属性可进行古代文学地理研究，辅助学者分析文历代文学家的地理分布。此外，利用本典籍知识图谱的相关古籍文献成书年代等信息，可以进一步考察中国古代学术的发展沿革情况。

典籍知识图谱是对古籍文献深层次开发与利用的一次尝试，对古籍文献目录知识服务的提升具有重要意义，典籍知识图谱可以对这些信息资源进行语义标注和链接，建立以知识为中心的资源语义集成服务。

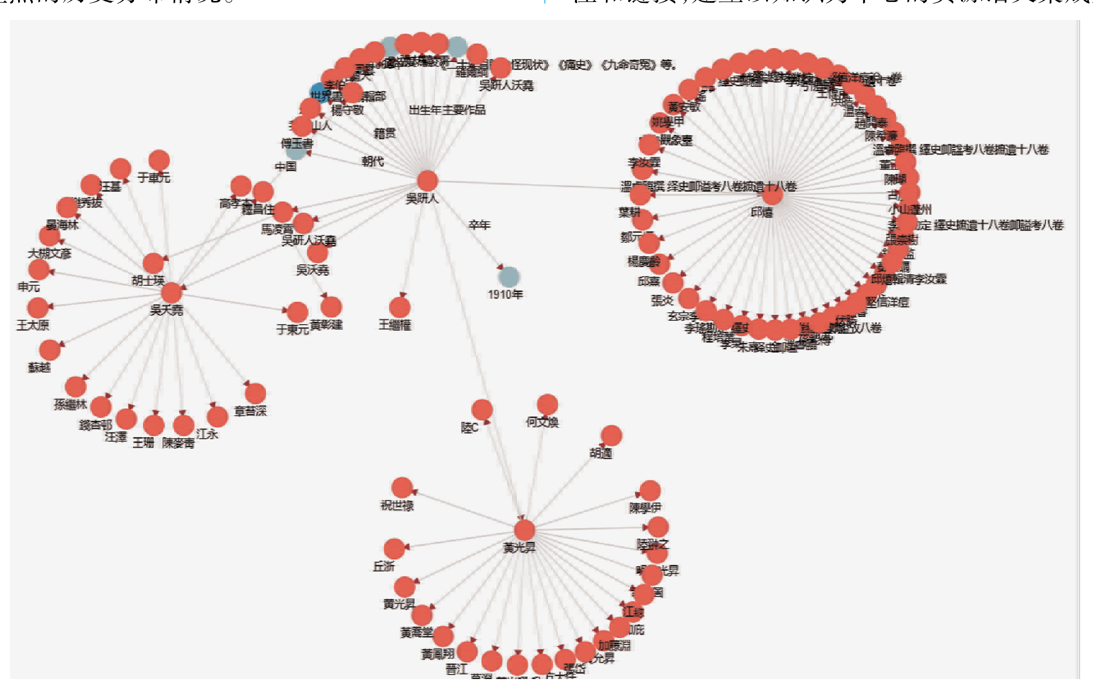


图 10 典籍合作者网络

6 结语

知识图谱是知识工程领域的一个最佳实践[28]，其基于图的结构更有利于典籍知识的表示及各知识单元的关联，以便于 典籍的存储、检索和知识服务。本研究通过古籍实体、典籍编撰者、古籍版本、地名这 4 类节点及其之间的关系构建了一个庞大的中国历代存世典籍知识图谱[29]，由节点、属性及边等形成了一个立体、多维、多用途的古籍知识关联网络，大大增强了古籍知识服务功能。

本典籍知识图谱基本由计算机完成典籍数据的获取、自动标注和切分，并在此基础上完成信息抽取工作和数据语义规范，由于缺乏足够的人工数据审校，计算机处理数据的过程中难免存在一些问题，数据的质量有待后续的提高，一些编撰者属性数据也有待补充完善。同时，典籍知识图谱的研究深度和高度有待进一步的研究与探索，比如典籍知识图谱的智能问答、知识推理等功能有待进一步深入研究与开发。

参考文献：

[1] 高路明. 古籍目录及其功用[J]. 文史知识，1981(5)：105 - 108.
[2] 鞠明库. 古籍数字化与传统文献学[J]. 清华大学学报(哲学社会科学版)，2011，26(5)：154 - 158，161.
[3] 杨清虎. 数字文献学的概念与问题[J]. 黑龙江史志，2013(13)：203.
[4] 于曼玲，余灼华. 用电子计算机编制古籍索引的体会[J]. 中山大学学报(哲学社会科学版)，1988(4)：95 - 96.
[5] 林仲湘. 编制《古今图书集成索引》的实践和理论[J]. 广西大学学报(哲学社会科学版)，1994(2)：94 - 102.
[6] 张琪玉. 古籍索引的一个范例——介绍《古今图书集成》电子版的索引数据库[J]. 图书馆杂志，2000(5)：48 - 49.
[7] 包菊香. 古籍目录索引的自动编制——以“中华古籍索引库”为例[J]. 中国索引，2013，11(1)：25 - 29.
[8] 何琳，曹玲. 农业古籍本体的构建及其检索机制研究[J]. 现代图书情报技术，2006(12)：37 - 39，53.
[9] 罗晨光，山川，王珊. 基于本体的古籍知识库建设初探[J]. 现代图书情报技术，2007(4)：8 - 11.
[10] “中国历代典籍总目分析系统”通过国家级技术鉴定[EB/OL]. [2020 - 03 - 21]. http：//pkunews. pku. edu. cn/xwzh/2009-11/02/content_161068. htm.
[11] 于生洪. 让书写在古籍里的文字活起来——中华书局古籍知识服务探索与实践[EB/OL]. [2020 - 05 - 21]. https：//news. artron. net/20180522/n1004873. html.
[12] 宋登汉，周迪，李明杰. 基于 RDA 的中国古籍版本资源描述设计(一)[J]. 图书馆，2010(4)：51 - 53.
[13] 宋登汉，周迪，李明杰. 基于 RDA 的中国古籍版本资源描述设计(二)[J]. 图书馆，2010(5)：49 - 52.
[14] 邓仲华，黄鑫，陆颖隽，等. 论中文古籍版本本体库的构建[J]. 图书情报知识，2014(4)：80 - 87，93.
[15] 夏翠娟，林海青，刘炜. 面向循证实践的中文古籍数据模型研究与设计[J]. 中国图书馆学报，2017，43(6)：16 - 34.
[16] 范佳. “数字人文”内涵与古籍数字化的深度开发[J]. 图书馆学研究，2013(3)：29 - 32.
[17] 徐清，石向实，王唯. 古籍数字化资源的深度开发[J]. 图书情报工作，2007，51(3)：95 - 97，79.
[18] 傅荣贤. 论章学诚“辨章学术，考镜源流”理念的本质[J]. 大学图书馆学报，2016，34(2)：111 - 117.
[19] 鲁欣. 从“辨章学术，考镜源流”看中国古典目录学之功用[J]. 江西图书馆学刊，2008(1)：9 - 11.
[20] 王国强. “辨章学术，考镜源流”之评判——中国古典目录学价值重估[J]. 郑州大学学报(哲学社会科学版)，1991(3)：77 - 81.
[21] 赵涛. 古籍目录史部学术源流与古代史学嬗变的历史路向[J]. 西北大学学报(哲学社会科学版)，2015，45(2)：26 - 32.
[22] 曾大兴. 中国历代文学家之地理分布[M]. 北京：商务印书馆，2013.
[23] 曹鑫. “一站式古籍目录检索系统”乃大势所趋[N]. 新华书目报，2017 - 02 - 17(12).
[24] 李涓子，侯磊. 知识图谱研究综述[J]. 山西大学学报(自然科学版)，2017，47(3)：454 - 459.
[25] 刘峤，李杨，段宏，等. 知识图谱构建技术综述[J]. 计算机研究与发展，2016，53(3)：582 - 600.
[26] 中华古籍总目编目规则[EB/OL]. [2020 - 03 - 09]. https：//max. book118. com/html/2017/0916/134109490. shtm.
[27] 陈涛，刘炜，单蓉蓉，等. 知识图谱在数字人文中的应用研究[J]. 中国图书馆学报，2019，45(6)：34 - 49.
[28] 赵一鸣. 知识图谱是一种知识组织系统吗？[J]. 图书情报知识，2017(5)：2.
[29] 中国古籍基础数据分析平台[EB/OL]. [2020 - 11 - 09]. http：//121. 201. 35. 124：88.

作者贡献说明：

欧阳剑：总体框架及功能设计，系统开发，论文构思及撰写；
梁珠芳：数据整理、校正等；
任树怀：论文修改。

欧阳剑，梁珠芳，任树怀. 大规模中国历代存世典籍知识图谱构建研究[J]. 图书情报工作,2021,65(5):126 - 135.

Research on the Construction of Knowledge Graph of Large-scale Chinese Ancient Books

Ouyang Jian[1,2] Liang Zhufang[3] Ren Shuhuai[1]

[1] Shanghai International Studies University Library, Shanghai 201620

[2] School of Journalism and Communication, Shanghai International Studies University, Shanghai 201620

[3] School of Management, Guangxi University for Nationalities, Nanning 530006

Abstract: [**Purpose/significance**] The establishment of a digital catalog is the need to protect and promote the Chinese civilization, and it also caters to the needs of new documentation and researchers. Chinese classics have been preserved throughout the ages. The construction of the knowledge graph provides a one-stop platform for researchers to dig out the hidden knowledge behind the massive bibliographic data of ancient books, which greatly enhances the knowledge service function of ancient books. The large-scale knowledge graph of ancient books is also an important foundation of machine intelligence. [**Method/process**] This research used knowledge graph technology to organize the knowledge of ancient Chinese classics, constructed a framework model of classics knowledge graph from three parts: demand layer, model layer, and application layer. Through man-machine collaboration, the data extraction of classics and multi-source data fusion, organize the data, analyze and define the entity types, attributes of the classic knowledge graph and the entity relationships, types of the classic knowledge graph. [**Result/conclusion**] It has realized the construction of the knowledge map of ancient books, including 649 549 kinds of ancient book entities, 221 783 persons in charge of ancient books, 1 498 383 versions of ancient books, 13 960 nodes of place names, and has formed a three-dimensional, multi-dimensional and multi-purpose knowledge association network of ancient books.

Keywords: ancient books knowledge organization knowledge graph humanities research digital humanities